Abstract:

This paper aims to document the implementations for the note-level singing voice transcription assignment for the NUS module CS4347 - Sound and Music Computing.

The last part of this article provides a summary of the improvement proposal of the current system.

Introduction:

• Singing data preprocessing

• Model training using processed singing data

• Transcription evaluation

Each steps will be explained in details in the following sessions.

Singing Data Preprocessing:

You are provided with a singing dataset with 100 pop songs and annotations of

each song, where 80 songs are for training, 10 songs are for validation, and 10

songs are for testing. Each song is a mixture of vocal and background music.

Before transcribing the singing voice, you first need to process the raw audios

in the given dataset. A simple and common strategy is taken and detailed in

the following steps:

1) Load the audio data and re-sample them with a sample rate = 44100 Hz

2) Compute Constant-Q Transform of the audio data

3) Segment all samples into frames with 1024 samples as one frame where

the time of the i-th frame is i ∗ 1024

44100